# **THÔNG TIN CHUNG CỦA NHÓM**

* Link YouTube video của báo cáo (tối đa 5 phút):   
  *(ví dụ: https://www.youtube.com/watch?v=AWq7uw-36Ng)*
* Link slides (dạng .pdf đặt trên Github của nhóm):   
  *(ví dụ: https://github.com/mynameuit/CS2205.xxx/TenDeTai.pdf)*
* *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
* *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*
* *Lớp Cao học, mỗi nhóm một thành viên*

| * Họ và Tên: Cao Quyết Chiến * MSSV: 240101005 | * Lớp: CS2205.NOV2024 * Tự đánh giá (điểm tổng kết môn): 8.5/10 * Số buổi vắng: 1 * Số câu hỏi QT cá nhân: 3 * Số câu hỏi QT của cả nhóm: 0 * Link Github: [Github](https://github.com/cqchien/Learnable_Regions_Clone/tree/d97987e5fac03609551ae6c36e14cdcfc52bb0e4/documents) |
| --- | --- |

# **ĐỀ CƯƠNG NGHIÊN CỨU**

| **TÊN ĐỀ TÀI (IN HOA)**  PHƯƠNG PHÁP CHỈNH SỬA HÌNH ẢNH DỰA TRÊN VĂN BẢN VỚI CẢI TIẾN VÙNG CHỈNH SỬA DỰA TRÊN PATCHES VÀ TINH CHỈNH ĐA CẤP |
| --- |
| **TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)**  TEXT-BASED IMAGE EDITING METHOD WITH IMPROVED PATCH-BASED REGIONS AND MULTI-SCALE REFINEMENT |
| **TÓM TẮT** *(Tối đa 400 từ)*  Text-driven image editing là một phương pháp mạnh mẽ cho việc chỉnh sửa hình ảnh trực quan, tận dụng các mô hình vision-language quy mô lớn. Tuy nhiên, các phương pháp hiện tại thường gặp khó khăn trong việc định vị chính xác và chỉnh sửa các vùng cụ thể, đặc biệt với các prompts phức tạp hoặc yêu cầu chỉnh sửa chi tiết. Dựa trên framework trong bài báo Text-Driven Image Editing via Learnable Regions, chúng tôi đề xuất cải tiến bằng cách tích hợp patch-based region optimization và multi-scale refinement. Phương pháp patch-based chia hình ảnh thành các vùng nhỏ, có thể học được, giúp xác định chính xác các khu vực cần chỉnh sửa mà không làm ảnh hưởng đến các vùng khác. Bên cạnh đó, multi-scale refinement đảm bảo quá trình chỉnh sửa được hòa trộn mượt mà vào toàn bộ hình ảnh, duy trì tính chân thực và sự phù hợp với ngữ cảnh. Mục tiêu của đề xuất là khắc phục các hạn chế của các phương pháp hiện tại, như định vị không chính xác hoặc xuất hiện artifacts trong kết quả chỉnh sửa, đồng thời nâng cao chất lượng và tính ứng dụng của hệ thống text-driven image editing. Các đánh giá thực nghiệm sẽ tập trung vào chất lượng chỉnh sửa cục bộ, mức độ phù hợp với prompts, và tính chân thực tổng thể của hình ảnh, qua đó chứng minh hiệu quả của phương pháp cải tiến. |
| **GIỚI THIỆU** *(Tối đa 1 trang A4)*  Text-driven image editing là một phương pháp mạnh mẽ cho phép người dùng chỉnh sửa hình ảnh trực quan thông qua các câu lệnh ngôn ngữ tự nhiên, tận dụng khả năng của các mô hình vision-language quy mô lớn. Phương pháp này đã mở ra nhiều ứng dụng tiềm năng trong các ngành công nghiệp sáng tạo, thiết kế và sản xuất nội dung nhờ tính đơn giản và trực quan. Tuy nhiên, các phương pháp hiện tại vẫn gặp nhiều hạn chế, đặc biệt trong việc định vị chính xác các vùng cần chỉnh sửa và đảm bảo sự tự nhiên trong kết quả đầu ra, nhất là khi xử lý các prompts phức tạp hoặc yêu cầu chỉnh sửa chi tiết. Những vấn đề này thường dẫn đến các kết quả chỉnh sửa không chính xác, xuất hiện artifacts hoặc thiếu sự hòa trộn mượt mà với ngữ cảnh tổng thể của hình ảnh.  Dựa trên framework từ bài báo Text-Driven Image Editing via Learnable Regions, nghiên cứu này đề xuất cải tiến bằng cách tích hợp hai kỹ thuật chính: patch-based region optimization và multi-scale refinement. Phương pháp patch-based chia hình ảnh thành các vùng nhỏ (patches) có thể học được, giúp định vị chính xác các khu vực cần chỉnh sửa mà không làm ảnh hưởng đến các vùng khác. Trong khi đó, multi-scale refinement đảm bảo các chỉnh sửa được hòa trộn mượt mà vào tổng thể hình ảnh, duy trì tính chân thực và phù hợp với ngữ cảnh. Hai kỹ thuật này không chỉ khắc phục những hạn chế của các phương pháp hiện tại mà còn nâng cao chất lượng và tính ứng dụng của hệ thống text-driven image editing.  Hệ thống được đề xuất sẽ nhận đầu vào là một hình ảnh và một câu lệnh văn bản, và trả về hình ảnh đã chỉnh sửa với chất lượng cao, đáp ứng đúng yêu cầu của người dùng. Các đánh giá thực nghiệm sẽ tập trung vào ba tiêu chí chính: độ chính xác của chỉnh sửa cục bộ, mức độ phù hợp với prompts và tính chân thực tổng thể của hình ảnh. Qua đó, nghiên cứu kỳ vọng sẽ chứng minh hiệu quả của phương pháp cải tiến, đồng thời đóng góp vào sự phát triển của lĩnh vực text-driven image editing, đáp ứng tốt hơn các yêu cầu thực tế. |
| **MỤC TIÊU** *(Viết trong vòng 3 mục tiêu)*   * Tích hợp kỹ thuật patch-based region optimization để tăng độ chính xác vùng chỉnh sửa: Xây dựng mô-đun chia hình ảnh thành các vùng nhỏ (patches) có khả năng học được, giúp định vị chính xác khu vực cần chỉnh sửa theo yêu cầu của prompt. Điều này đảm bảo rằng các chỉnh sửa chỉ tác động đến vùng mục tiêu mà không làm ảnh hưởng đến các khu vực không liên quan trong hình ảnh. * Áp dụng kỹ thuật multi-scale refinement để duy trì tính tự nhiên và hòa trộn mượt mà: Phát triển phương pháp refinement đa cấp độ để đảm bảo các chỉnh sửa được hòa trộn một cách tự nhiên vào toàn bộ hình ảnh. * Đánh giá hiệu quả hệ thống qua các tiêu chí thực nghiệm quan trọng: Độ chính xác chỉnh sửa cục bộ, mức độ phù hợp với prompt, tính chân thực tổng thể |
| **NỘI DUNG VÀ PHƯƠNG PHÁP**  Hệ thống chỉnh sửa hình ảnh dựa trên văn bản (text-driven image editing) được phát triển thông qua ba nội dung chính.  **1. Patch-based Region Optimization**  Một trong những thách thức lớn của chỉnh sửa hình ảnh dựa trên văn bản là định vị chính xác khu vực cần chỉnh sửa trên hình ảnh. Để giải quyết vấn đề này, nghiên cứu áp dụng kỹ thuật tối ưu hóa vùng chỉnh sửa dựa trên các vùng nhỏ (patches). Trước tiên, hệ thống sử dụng các mô hình vision-language như CLIP để ánh xạ câu lệnh ngôn ngữ tự nhiên (prompt) vào không gian hình ảnh, qua đó xác định mối liên kết giữa các vùng hình ảnh và ý nghĩa của prompt. Sau đó, hình ảnh được chia thành các vùng nhỏ (patches), mỗi vùng sẽ được phân tích mức độ liên quan đến prompt. Các vùng liên quan sẽ được tối ưu hóa bằng cơ chế attention, đảm bảo rằng chỉ khu vực cần chỉnh sửa bị tác động, trong khi các vùng không liên quan được giữ nguyên. Phương pháp này giúp nâng cao độ chính xác của vùng chỉnh sửa, giảm thiểu các thay đổi không mong muốn trên hình ảnh.  **2. Multi-scale Refinement**  Để đảm bảo tính tự nhiên và hòa trộn mượt mà giữa các chỉnh sửa và ngữ cảnh tổng thể của hình ảnh, nghiên cứu tích hợp kỹ thuật tinh chỉnh đa cấp độ. Phương pháp này phân tích hình ảnh ở nhiều cấp độ khác nhau, từ các đặc trưng cấp thấp như màu sắc và kết cấu, đến các đặc trưng cấp cao như ánh sáng và bố cục tổng thể. Một mô-đun tinh chỉnh chuyên biệt được phát triển để điều chỉnh các chỉnh sửa ở từng cấp độ, đảm bảo sự nhất quán về ánh sáng, màu sắc và cấu trúc tổng thể của hình ảnh. Ngoài ra, mạng GAN (Generative Adversarial Network) được tích hợp để kiểm tra và cải thiện tính tự nhiên của hình ảnh sau chỉnh sửa. Generator sẽ tạo ra hình ảnh chỉnh sửa, trong khi Discriminator đánh giá mức độ chân thực và tự nhiên, giúp hệ thống đạt được kết quả chỉnh sửa mượt mà và không xuất hiện artifacts.  **3. Đánh giá hiệu quả hệ thống thông qua các bài kiểm tra thực nghiệm**  Để kiểm tra chất lượng và hiệu quả của hệ thống, nghiên cứu thiết kế một quy trình đánh giá toàn diện, bao gồm cả phương pháp tự động và thủ công. Quy trình đánh giá tự động sử dụng các chỉ số như SSIM (Structural Similarity Index) để đo độ tương đồng giữa hình ảnh chỉnh sửa và hình ảnh mong muốn. Đồng thời, đánh giá thủ công được thực hiện thông qua khảo sát người dùng, trong đó người tham gia sẽ đánh giá mức độ phù hợp của chỉnh sửa với prompt và tính tự nhiên của hình ảnh trên thang điểm Likert. Cuối cùng, hệ thống được so sánh với các phương pháp hiện có như DALL-E hoặc Stable Diffusion để làm nổi bật các cải tiến về độ chính xác và chất lượng chỉnh sửa.  Ba nội dung trên được triển khai theo trình tự logic, từ việc phát triển các kỹ thuật định vị và chỉnh sửa (patch-based region optimization và multi-scale refinement) đến việc kiểm tra hiệu quả qua các bài đánh giá thực nghiệm. Các phương pháp này không chỉ đảm bảo tính khả thi của nghiên cứu mà còn mở ra tiềm năng ứng dụng rộng rãi trong các lĩnh vực như thiết kế sáng tạo, sản xuất nội dung và truyền thông. |
| **KẾT QUẢ MONG ĐỢI**  *Nghiên cứu dự kiến xây dựng một hệ thống chỉnh sửa hình ảnh dựa trên văn bản (text-driven image editing) với khả năng định vị chính xác khu vực cần chỉnh sửa và thực hiện các thay đổi một cách tự nhiên, chân thực. Người dùng có thể nhập các câu lệnh ngôn ngữ tự nhiên để yêu cầu chỉnh sửa hình ảnh, từ thay đổi màu sắc, ánh sáng đến thêm, xóa hoặc thay thế đối tượng. Nhờ áp dụng kỹ thuật tối ưu hóa vùng chỉnh sửa (patch-based region optimization) và tinh chỉnh đa cấp độ (multi-scale refinement), hệ thống đảm bảo các thay đổi chỉ tác động đến vùng mục tiêu, đồng thời duy trì tính nhất quán về ánh sáng, màu sắc và cấu trúc tổng thể của hình ảnh.*  *Hệ thống sẽ được triển khai dưới dạng ứng dụng web với giao diện thân thiện, cho phép tải lên hình ảnh, nhập lệnh chỉnh sửa, xem trước kết quả và tải xuống hình ảnh đã chỉnh sửa. Hệ thống hỗ trợ đa dạng loại hình ảnh và tình huống chỉnh sửa, từ các yêu cầu cơ bản đến sáng tạo phức tạp, đồng thời tích hợp cơ chế đánh giá tự động để cải thiện chất lượng theo thời gian. Kết quả nghiên cứu không chỉ cung cấp một công cụ chỉnh sửa hình ảnh hiện đại mà còn mở rộng tiềm năng ứng dụng trong thiết kế sáng tạo, sản xuất nội dung số, truyền thông, quảng cáo và nhu cầu cá nhân, mang lại giá trị thực tiễn cao cho người dùng.* |
| **TÀI LIỆU THAM KHẢO**  [1] Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Lu Jiang, Ming-Hsuan Yang:  Text-Driven Image Editing via Learnable Regions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1–13.  [2] Patrick Esser, Robin Rombach, Björn Ommer:  Taming Transformers for High-Resolution Image Synthesis. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12873–12883.  [3] Jonathan Ho, Ajay Jain, Pieter Abbeel: Denoising Diffusion Probabilistic Models. Advances in Neural Information Processing Systems (NeurIPS), 2020, pp. 6840–6851.  [4] Ruiqi Gao, Xiaoyong Shen, Jiaya Jia: MaskGIT: Masked Generative Image Transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11315–11325.  [5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever:  Learning Transferable Visual Models From Natural Language Supervision. Proceedings of the 38th International Conference on Machine Learning (ICML), 2021, pp. 8748–8763.  [6] A. A. Efros, W. T. Freeman: Image Quilting for Texture Synthesis and Transfer. Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), 2001, pp. 341–346.  [7] Connelly Barnes, Eli Shechtman, Adam Finkelstein, Dan B. Goldman:  PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing. ACM Transactions on Graphics (TOG), Volume 28, Issue 3, 2009, Article No. 24.  [8] Jian Sun, Nan-Ning Zheng, Heung-Yeung Shum: Image Completion with Structure Propagation. ACM Transactions on Graphics (TOG), Volume 22, Issue 3, 2003, pp. 861–868.  [9] Olaf Ronneberger, Philipp Fischer, Thomas Brox: U-Net: Convolutional Networks for Biomedical Image Segmentation. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234–241.  [10] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, Ming-Hsuan Yang:  Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 624–632.  [11] Guang-Hai Liu, Jing-Yu Yang:  Exploiting Color Volume and Color Difference for Salient Region Optimization. IEEE Transactions on Image Processing, Volume 28, 2019, pp. 56–68.  [12] Mingliang Zhou:  Manifold and Patch-based Unsupervised Deep Metric Learning for Fine-grained Image Retrieval. Proceedings of the International Conference on Multimedia Retrieval (ICMR), 2020, pp. 1–9.  [13] Pan Gao:  Patch-Based Deep Autoencoder for Point Cloud Geometry Compression. Proceedings of the ACM Multimedia Asia (MMAsia), 2022, pp. 1–5.  [14] Zhuoyuan Chen, et al.:  Learning Multi-scale Features for Image Segmentation Using Deep Convolutional Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2020, pp. 1–15.  [15] Xiang Li, et al.:  Multi-scale Attention Network for Image Super-Resolution. IEEE Access, Volume 9, 2021, pp. 1–10.  [16] Jianbo Shi, Jitendra Malik:  Normalized Cuts and Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Volume 22, Issue 8, 2000, pp. 888–905.  [17] Alex Kendall, Yarin Gal:  What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5574–5584. |